MATERIALS AND METHODS FOR IDENTIFYING AND ANALYZING INTERMEDIATE TANDEM REPEAT DNA MARKERS

CROSS-REFERENCE TO RELATED APPLICATIONS

Not Applicable.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

This invention was made with support from the United States Government, under Small Business Innovation Research Grant Numbers 1-43-MH5294-01 and 1-43-MH5294-02, awarded by the National Institutes of Health. The United States Government has certain rights in the invention.

FIELD OF THE INVENTION

The present invention is generally directed to the identification and analysis of genetic markers in a genomic system. The present invention is more specifically directed to the identification of loci in DNA, particularly in genomic DNA, containing length polymorphisms due to variations in the number of intermediate (5 to 7 base) sequence repeats. The present invention is also directed to the detection of such polymorphic loci. The invention is directed, furthermore, to methods of identifying and distinguishing individuals based primarily on differences in size of the products of amplifying genomic DNA at such a locus, wherein the number of intermediate tandem repeat sequences vary from one individual to another.

BACKGROUND OF THE INVENTION

DNA typing is commonly used to identify the parentage of human children, and to confirm the lineage of horses, dogs, and other prize animals. DNA typing is also commonly employed to identify the source of blood, saliva, semen, and other tissue found at a crime scene. DNA typing methods in use today are designed to detect and analyze differences in the length and/or sequence of one or more regions of DNA known to appear in at least two different forms in a population. DNA typing is

also employed in clinical settings to determine success or failure of bone marrow transplantation and presence of particular cancerous tissues. Such length and/or sequence variation is referred to as "polymorphism." Any region (i.e. "locus") of DNA in which such a variation occurs is referred to as a "polymorphic locus." Most DNA typing techniques employ at least one "marker" containing the at least one such polymorphic locus. Each individual marker contains a single allele of genomic DNA ultimately derived from a single individual in a population. The methods and materials of the present invention are all designed for use in the detection of a particular class of polymorphisms in DNA characterized primarily by variation in length.

Genetic markers which are sufficiently polymorphic with respect to length or sequence have long been sought for use in identity applications, such as paternity testing and identification of tissue samples collected for forensic analysis. The discovery and development of such markers and methods for analyzing such markers have gone through several phases of development over the last several years. In recent years, the discovery and development of polymorphic short tandem repeats (STRs) as genetic markers has stimulated progress in the development of linkage maps, the identification and characterization of diseased genes, and the simplification and precision of DNA typing. The term "short tandem repeat" or "STR" as used herein refers to all sequences between two and seven nucleotides long which are repeated perfectly, or nearly perfectly in tandem within the genomic DNA of any organism. See, for example, the definition of "short tandem repeat" applied to human genomic DNA in U.S. Pat. No. 5,364,759, column 4, line 58 *et seq.*

The first identified DNA variant markers were simple base substitutions, i.e. simple sequence polymorphisms, which were most often detected by Southern hybridization assays. For examples of references describing the identification of such markers, designed to be used to analyze restriction endonuclease-digested DNA with radioactive probes, see: Southern, E. M. (1975), J. Mol. Biol. 98(3):503-507; Schumm, et al. (1988), *American Journal of Human Genetics* 42:143-159; and Wyman, A. and White, R. (1980) *Proc. Natl. Acad. Sci, U.S.A.* 77:6754-6758.

The next generation of markers were size variants, i.e. length polymorphisms, specifically "variable number of tandem repeat" (VNTR) markers (Nakamura Y., et al. (1987), *Science* 235: 1616-1622; and U.S. Pat. No. 4,963,663 issued to White et

al. (1990); U.S. Pat. No. 5,411,859 continuation of 4,963,663 issued to White et al. (1995)) and "minisatellite" markers (Jeffreys et al. (1985a), *Nature* 314:67-73; Jeffreys et al. (1985b) *Nature* 316:76-79., U.S. Pat. No. 5,175,082 for an invention by Jeffreys). Both VNTR and minisatellite markers, contain regions of nearly identical sequences repeated in tandem fashion. The core repeat sequence is 10 to 70 bases in length, with shorter core repeat sequences referred to as "minisatellite" repeats and longer repeats referred to as VNTRs. Different individuals in a human population contain different numbers of these repeats. These markers are more highly polymorphic than base substitution polymorphisms, sometimes displaying up to forty or more alleles at a single genetic locus. However, the tedious process of restriction enzyme digestion and subsequent Southern hybridization analysis are still required to detect and analyze most such markers.

The next advance involved the joining of the polymerase chain reaction (PCR) (U.S. Pat. No. 4,683,202 by Mullis, K.B.) technology with the analysis of VNTR loci (Kasai K, et al. (1990) Journal Forensic Science 35(5):1196-1200). Amplifiable VNTR loci were discovered, which could be detected without the need for Southern transfer. The amplified products are separated through agarose or polyacrylamide gels and detected by incorporation of radioactivity during the amplification or by post-staining with silver or ethidium bromide. However, PCR can only be used to amplify relatively small DNA segments reliably, i.e. only reliably amplifying DNA segments under 3,000 bases in length Ponce, M & Micol, L. (1992) NAR 20(3):623; Decorte R, et al. (1990) DNA Cell Biol. 9(6):461-469). Consequently, very few amplifiable VNTRs have been developed, making them, as a class, impractical for linkage mapping.

With the recent development of polymorphic markers with polymorphic dinucleotide repeats (Litt and Luty (1989) *Am J. Hum Genet* 3(4):599-605; Tautz, D (1989) *NAR* 17:6463-6471; Weber and May (1989) *Am J Hum Genet* 44:388-396; German Pat. No. DE 38 34 636 C2, inventor Tautz, D; U.S. Pat. No. 5,582,979 filed by Weber, L.) and with polymorphic short tandem repeats (STR) (Edwards, A., et al. (1991) *Am. J. Hum. Genet.* 49: 746-756.; Hammond, H.A., et al. (1994) *Am. J. Hum. Genet.* 55: 175-189; Fregeau, C.J.; and Fourney, R.M. (1993) *BioTechniques* 15(1): 100-119.; Schumm, J.W. et al. (1994) in The Fourth International Symposium on Human Identification 1993, pp. 177-187; and U.S. Pat. No. 5,364,759 by Caskey et

al.; German Pat. No. DE 38 34 636 C2 by Tautz, D.) many of the deficiencies of previous methods have been overcome. The two types of markers, those containing dinucleotide or STR repeats (which by definition include 2-7 bp repeats), are generally referred to as "microsatellite" markers. Often considered to be the best available markers, the microsatellite loci are similar to amplifiable VNTRs, in that their alleles may be differentiated based on length variation. However, unlike VNTRs, these loci contain perfect or imperfect repeat sequences two, three, four, or rarely, five bases long. They display from just a few alleles to more than forty at a single locus. Amplification protocols can be designed to produce small products, generally from 60 to 400 base pairs long, and alleles from each locus are often contained within a range of less than 50 bp. This allows simultaneous electrophoretic analysis of several systems on the same gel by careful design of PCR primers such that all potential amplification products from an individual system do not overlap the range of alleles of other systems in the same gel.

Three significant drawbacks relate to the use of microsatellite loci. First, the presence of stutter artifacts, that is, one or more minor fragments in additional to the major fragment representing each allele, is often seen following amplification. This deficiency is much more severely displayed with dinucleotide repeat loci than with tri- or tetranucleotide repeat markers (Edwards et al., 1991. Am J Hum Genet 49;746-756; Edwards et al., 1992. Genomics 12:241-253; Weber & May, 1989. Am J Hum Genet 44:388-396). The presence of these artifacts, presumed to result from a DNA polymerase-related phenomenon called repeat slippage (Levinson & Gutman, 1987. Mol. Biol. Evol. 4(3):203-221; Schlotterer & Tautz, 1992. NAR 20:211-215), complicates the interpretation of allelic content of the loci. While complicating all interpretations, the presence of major and minor fragments to represent each allele especially limits the usefulness of these markers in forensic analysis which often require determination of whether more than one source of DNA sample is present. Many of the markers described in this work represent a new class of markers which produce significantly less stutter artifact than known markers.

A second drawback to current STR and microsatellite marker systems relates to the difficulty in separating multiple loci in a single gel. This occurs because there is spacial compression of fragments of different size in the upper regions of the gels most commonly used for separation of DNA fragments by those skilled in the art.

Development of the markers described in this work, based on larger repeat units, extends the useful range within these gels, allowing simultaneous analysis of more loci.

A third drawback is that, prior to the invention disclosed herein, only a few DNA loci of human genomic DNA had been described in the literature, with length polymorphisms based on variations in a number of five to seven base repeats at each such locus. See, e.g. Edwards et al. (1991) *Nucleic Acids Res.* 19:4791; Chen et al. (1993) *Genomics* 15(3): 621-5; Harada et al. (1994) *Am. J. Hum. Genet.* 55: 175-189; Comings et al. (1995), *Genomics* 29(2):390-6; and Utah Marker Development Group (1995), *Am. J. Genet.* 57:619-628. In 1995, Jurka and Pethiyagoda published an article describing a study in which they had used the GenBank database to determine the relative abundance and variability of pentameric and hexameric tandem repeats in the primate genome (Jurka and Pethiyagoda (1995) *J. Mol. Evol.* 40:120-126). However, variability was only indirectly estimated, and polymorphism levels at individual loci were not demonstrated. *Id.* We have developed materials and methods for identifying and analyzing DNA loci which contain highly polymorphic repeats of five to seven base repeats.

The materials and methods of the present method are designed for use in identifying and analyzing particular polymorphic loci of DNA of various types, including single-stranded and double-stranded DNA from a variety of different sources. The present invention represents a significant improvement over existing technology, bringing increased power and precision to DNA profiling for linkage analysis, criminal justice, paternity testing, and other forensic and medical uses.

BRIEF SUMMARY OF THE INVENTION

It is, therefore, an object of the present invention to provide materials and methods for the identification and analysis of DNA loci with intermediate tandem repeat sequences, wherein an "intermediate tandem repeat sequence" is a region of DNA which contains at least one repeat unit consisting of a sequence of five (5), six (6), or seven (7) bases repeated in tandem at least two (2) times.

Another object of the present invention is to provide materials and methods for identifying intermediate tandem repeat DNA markers, which produce fewer artifacts when used to analyze or detect one or more loci of a DNA sample containing

an intermediate tandem repeat. The methods and materials of the present invention are preferably used to identify and analyze loci of genomic DNA, each of which contains a polymorphic intermediate tandem repeat sequence. The materials of this invention include oligonucleotide primers and DNA markers to such loci of human genomic DNA. Intermediate tandem repeat loci detected using methods of the present invention exhibit fewer artifacts than do many known loci detected using similar methods, including short STR's (i.e. tandem repeats of a two, three or four base DNA sequence).

A particular object of the present invention is to provide a method and materials for the analysis of individual polymorphic genetic loci based primarily on length variation due primarily to differences in the number of nucleic acid repeat units in a region of intermediate nucleic acid tandem repeats. It is also a specific object of the present invention to provide a method, a kit, and primers for the detection and analysis of a polymorphic loci of genomic DNA, containing intermediate tandem repeat polymorphisms, including pentanucleotide tandem repeat polymorphisms.

One embodiment of the present invention consists of a method of isolating a fragment of DNA containing an intermediate tandem repeat sequence from genomic DNA, comprising: (a) providing a plurality of fragments of DNA, wherein at least one fragment contains an intermediate tandem repeat sequence; (b) providing a support means, e.g. a stationary support means, having associated therewith at least one oligonucleotide comprising a sequence of nucleotides which is complementary to a portion of the intermediate tandem repeat sequence; and (c) combining the plurality of fragments of DNA with the support means under conditions wherein the DNA fragment containing the intermediate repeat sequence and at least one other DNA fragment hybridizes to the support means.

An alternative embodiment of the invention is a method for detecting a polymorphic intermediate tandem repeat sequence having a low incidence of stutter artifacts in genomic DNA, comprising: (a) providing a sample of DNA having at least one target intermediate tandem repeat sequence, and (b) detecting the target intermediate tandem repeat sequence in the sample of DNA, wherein an average stutter artifact of no more than 1.1% is observed.

An additional embodiment of the invention is a method for detecting a target intermediate tandem repeat sequence in a DNA sample using at least one

oligonucleotide primer to amplify an intermediate tandem repeat sequence of interest (hereinafter, the "target intermediate tandem repeat sequence) in the sample DNA, wherein the oligonucleotide primer comprises a sequence which is complementary to and flanks a region of a DNA marker containing an intermediate tandem repeat sequence (hereinafter, the "template intermediate tandem repeat sequence") in the DNA marker sequence, wherein the DNA marker has a sequence selected from the group of sequences consisting of SEQ ID NO's: 1 through 43.

In another embodiment, the invention is a kit for the detection of at least one target intermediate tandem repeat sequence in a sample of DNA, the kit comprising a container which has at least one oligonucleotide primer for amplifying the at least one target intermediate tandem repeat sequence, wherein the oligonucleotide primer comprises a sequence of nucleotides which is complementary to and flanks a portion of a region of a double-stranded DNA marker containing a template intermediate tandem repeat sequence, wherein the DNA marker has a sequence selected from the group consisting of SEQ ID NO:'s 1 through 43.

In yet another embodiment, the invention is an oligonucleotide primer comprising a sequence complementary to a strand of a double-stranded DNA marker in a region of the marker flanking a template intermediate tandem repeat sequence, wherein the DNA marker has a sequence selected from the group consisting of: SEQ ID NO:'s 1 through 6, and SEQ ID NO:'s 28 through 33.

Each of the various embodiments of the present invention have specific use in the fields of human and other organism identification, forensic analysis, paternity determination, monitoring of bone marrow transplantation, linkage mapping, and detection of genetic diseases and cancers. The need to distinguish accurately between small amounts of tissue of different individuals is particularly acute in forensics applications, where many convictions (and acquittals) depend on DNA typing analysis, including the analysis of STR loci.

Further objects, features, and advantages of the invention will be apparent from the following best mode for carrying out the invention and the illustrative drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a flow diagram of a method of intermediate tandem repeat enrichment by filter hybridization.

FIG. 2 is an electropherogram of an S159 pentanucleotide repeat.

FIG. 3 is an electropherogram of a vWA tetranucleotide repeat.

FIG. 4 is an electropherogram of a G210 pentanucleotide repeat.

FIG. 5 is an electropherogram of a D5S818 tetranucleotide repeat.

FIG. 6 is a scatter plot of % stutter of the S159 pentanucleotide repeat.

FIG. 7 is a scatter plot of % stutter of the G210 pentanucleotide repeat.

FIG. 8 is a scatter plot of % stutter of the D5S818 tetranucleotide repeat.

FIG. 9 is a scatter plot of % stutter of the vWA tetranucleotide repeat.

FIG. 10 is a laser printed image of the results of fluorimager scan of fluorescent labeled amplified fragments of a S159 pentanucleotide repeat, after separation by gel electrophoresis.

FIG. 11 is a laser printed image of the results of fluorimager scan of fluorescent labeled amplified fragments of a G210 pentanucleotide repeat, after separation by gel electrophoresis.

The drawings and figures are not necessarily to scale and certain features of the invention may be exaggerated in scale or shown in schematic form in the interest of clarity and conciseness.

DETAILED DESCRIPTION OF THE INVENTION

It will be readily apparent to one skilled in the art that various substitutions and modifications may be made to the invention disclosed herein without departing from the scope and the spirit of the invention.

A. Definitions:

As used herein, the term "intermediate tandem repeat" or "ITR" refers to a region of a DNA sequence comprising a five to seven base sequence repeated in tandem at least two times. The term ITR also encompasses a region of DNA wherein more than a single five to seven base sequence is repeated in tandem or with intervening bases, provided that at least one of the sequences is repeated at least two times in tandem. Each sequence repeated at least once within an ITR is referred

to herein as a "repeat unit."

An "ITR polymorphism" refers an ITR in genomic DNA which varies in length from one chromosome to another in a population of individuals, due primarily to differences in the number of repeat units in the same region of each chromosome.

The intermediate tandem repeat sequences identified and analyzed according to the present invention can be divided into two general categories, perfect and imperfect. The term "perfect" ITR, as used herein, refers to a region of double-stranded DNA containing a single five to seven base repeat unit repeated in tandem at least two times, e.g. $(AAAAT)_{12}$. The term "imperfect" ITR, as used herein, refers to a region of DNA containing at least two tandem repeats of a perfect repeat unit and at least one repeat of an imperfect repeat unit, wherein the imperfect repeat unit consists of a DNA sequence which could result from one, two, or three base insertions, deletions, or substitutions in the sequence of the perfect repeat unit, e.g. $(AAAAT)_{12}(AAAAAT)_5AAT(AAATT)_4$. Every imperfect ITR sequence contains at least one perfect ITR sequence. Specifically, every ITR sequence, whether perfect or imperfect, includes at least one repeat unit sequence appearing at least two times in tandem, a repeat unit sequence which can be represented by formula (I):

$$(A_wG_xT_yC_z)_n \qquad (I)$$

wherein A, G, T, and C represent the nucleotides which can be in any order; w, x, y and z represent the number of each nucleotide in the sequence and range from 0 to 7 with the sum of w+x+y+z ranging between 5 and 7; and n represents the number of times the sequence is tandemly repeated and is at least 2.

"Pentanucleotide tandem repeat" refers to a subclass of the "intermediate tandem repeat" polymorphisms defined above. Unless specified otherwise, the term "pentanucleotide tandem repeat" encompasses perfect ITRs wherein the repeat unit is a five base sequence, and imperfect ITRs wherein at least one repeat unit is a five base repeat.

"DNA Marker" refers to a fragment of DNA which contains an ITR sequence such as a fragment of DNA containing an ITR sequence produced by amplifying a region of genomic DNA. Each individual marker contains a single allele of genomic DNA ultimately derived from a single individual in a population.

The term "locus" refers to a specific region of DNA. When used to describe a region of genomic DNA, "locus" refers to a particular position on a chromosome.

The same genomic locus appears at identical sites on each pair of homologous chromosomes for any individual in a population. The sequence of DNA at the same locus on each such chromosome, or at the same locus of DNA originating from the same such chromosome, is referred to as an "allele."

The term "polymorphism", as used herein refers to variations in the alleles at a locus seen in at least two chromosomes found in the genomic DNA of a population of individual organisms of the same species. The term "polymorphism" includes variations in the sequence of DNA obtained from the same locus of fragments of chromosomes cloned into other vehicles, such as DNA vectors or the chromosomal DNA of another organism.

As used herein, "ITR flanking sequence" refers to the nucleotide sequence adjacent to an ITR on a strand of DNA sequence containing an ITR. Sequences which include the ITR flanking sequence as a portion of their entire sequence are themselves flanking sequences.

The term "oligonucleotide primer" as used herein defines a molecule comprised of more than three deoxyribonucleotides or ribonucleotides. Although each primer sequence need not reflect the exact sequence of the template, the more closely the sequence reflects the complementarity to a template, the better the binding to the template. Its exact length and sequence will depend on many factors relating to the ultimate function and use of the oligonucleotide primer, including temperature, sequence of the primer, and use of the method. Each oligonucleotide primer of the present invention comprises a sequence of nucleic acids which is complementary to the sequence of a DNA marker flanking an ITR sequence. The oligonucleotide primers of the present invention are capable of acting as an initiation point for synthesis when placed under conditions which induce synthesis of a primer extension product complementary to a nucleic acid strand. The conditions can include the presence of nucleotides and an inducing agent, such as a DNA polymerase at a suitable temperature and pH. In the preferred embodiment, the primer is a single-stranded oligodeoxyribonuclotide of sufficient length to prime the synthesis of an extension product from a specific sequence in the presence of an inducing agent. Sensitivity and specificity of the oligonucleotide primers are determined by the primer length and uniqueness of sequence within a given sample of template DNA. In the present invention the oligonucleotide primers are usually

about greater than 15 bases and preferably about 20 to 40 bases in length.

The term "oligonucleotide primer pair" refers to a pair of primers, each comprising a sequence of deoxyribonucleotide or ribonucleotide bases complementary to opposite strands of double-stranded DNA flanking the same ITR. Each pair of oligonucleotide primers of the present invention is preferably selected to detect a single ITR. Although each primer sequence need not reflect the exact sequence of the template, the more closely the sequence reflects the complementarity to a template, the better the binding to the template.

The term "extension product" refers to the nucleotide sequence which is synthesized from the 3' end of the oligonucleotide primer and which is complementary to the strand to which the oligonucleotide is bound.

The term "oligonucleotide probe", as used herein, refers to a single-stranded molecule of DNA or RNA comprising a sequence which is complementary to a portion of a target sequence, such as the intermediate tandem repeat sequence of a DNA sample, wherein the portion of complementarity is of sufficient length to enable the probe to hybridize to the target sequence.

The term "stutter artifact", as used herein, refers to a particular type of artifact observed when detecting one or more molecules of target DNA, wherein the target DNA contains tandem repeats of the same repeat unit sequence, including the target intermediate tandem repeat sequences detected and analyzed according to the present invention. When a sample containing any such target DNA is detected after separation of all DNA in the sample by length, e.g. using gel electrophoresis, each molecule of target DNA produces a major signal (e.g. a major band on a gel); but, a minor signal can be detected proximate to each major signal. The minor signal is generally produced from the detection of DNA fragments which differ from the target DNA in length due to the addition or deletion of one or more repeat units from the target DNA sequence. Stutter artifacts have been attributed to slipped-strand mispairing during replication of DNA, both *in vivo* and *in vitro*. See, e.g. Levinson and Gutman (1987), *Mol. Biol. Evol,* 4(3):203-221; and Schlötterer and Tautz (1992), *Nucleic Acids Research* 20(2):211-215. Such artifacts are particularly apparent when DNA containing any such repeat sequence is amplified *in vitro*, using a method of amplification such as the polymerase chain reaction (PCR), as any minor fragment present in a sample or produced during polymerization is amplified along with the

major fragments.

The term "% stutter artifact" as used herein refers to a comparison of the amplitude of a minor (i.e. artifact) signal to the amplitude of a major (i.e. target) signal observed in a sample of DNA obtained from a single source, such as a single colony of bacteria or a single chromosome of genomic DNA. % stutter artifact can be determined on DNA which has not been amplified; but, is preferably determined after amplification of at least one target intermediate tandem repeat sequence. The term "average % stutter artifact" refers to an average of % stutter artifacts obtained from the measurements of % stutter artifact detected from a representative sample of at least twenty alleles in a population.

The term "genomic DNA" as used herein refers to any DNA ultimately derived from the DNA of a genome. The term includes, for example, cloned DNA in a heterologous organism, whole genomic DNA, and partial genomic DNA (e.g. the DNA of a single isolated chromosome).

The DNA detected or isolated according to the present invention can be single-stranded or double-stranded. For example, single-stranded DNA suitable for use in the present invention can be obtained from bacteriophage, bacteria, or fragments of genomic DNA. Double-stranded DNA suitable for use in the present invention can be obtained from any one of a number of different sources containing DNA with intermediate tandem repeat sequences, including phage libraries, cosmid libraries, and bacterial genomic or plasmid DNA, and DNA isolated from any eukaryotic organism, including human genomic DNA. The DNA is preferably obtained from human genomic DNA. Any one of a number of different sources of human genomic DNA can be used, including medical or forensic samples, such as blood, semen, vaginal swabs, tissue, hair, saliva, urine, and mixtures of bodily fluids. Such samples can be fresh, old, dried, and/or partially degraded. The samples can be collected from evidence at the scene of a crime.

B. Method of Isolating Polymorphic DNA Markers Containing an ITR:

One embodiment of the present invention is a method for isolating a fragment of DNA containing an ITR, using hybridization selection. The method comprises the steps of: (a) providing a plurality of fragments of DNA, wherein at least one DNA fragment contains an ITR; (b) providing a support means having at least one

oligonucleotide associated therewith, wherein the oligonucleotide includes a sequence of nucleotides which is complementary to a portion of the intermediate tandem repeat sequence; and (c) combining the plurality of fragments of DNA with the support means under conditions wherein DNA fragments, including any DNA fragments containing the ITR sequence, hybridize to the support means.

The plurality of fragments of DNA provided in step (a) of the method can be obtained by fragmenting any sample of DNA containing an ITR, but are preferably obtained by fragmenting genomic DNA. See, e.g. Current Protocols in Human Genetics (1994), Chapter 2: Development of Genetic Markers, Construction of Small-Insert Libraries from Genomic DNA, p. 2.2.1 *et seq.*, which is incorporated herein by reference. The most preferred method for preparing a plurality of fragments of DNA for use in step (a) is according to the steps comprising: fragmenting a sample of DNA, thereby producing a population DNA fragments wherein at least one DNA fragment contains the ITR; ligating a linker containing a priming sequence to at least one end of each DNA fragment in the population DNA fragments; and amplifying each linker ligated fragment using an oligonucleotide primer comprising a sequence which is complementary to the priming sequence. A different linker can be ligated to each end of each fragment. However, a single linker is preferably ligated to each end to enable amplification using a single oligonucleotide primer having a sequence which is complementary to the priming sequence of the linker. Linker ligation is preferably conducted in the presence of a ligase enzyme, such as T4 DNA ligase.

Any one of a number of different means can be used to produce the plurality of DNA fragments provided in step (a) of the method, including sonication or fragmentation with at least one restriction enzyme, although only double-stranded DNA can be fragmented with a restriction enzyme. When a restriction enzyme is used to fragment a sample of double-stranded DNA, it is preferably a restriction enzyme with a four base pair recognition sequence, which leaves single-stranded overhangs, and which does not cut the DNA sample within the ITR region of interest. Preferred restriction enzymes for use in fragmenting a double-stranded DNA sample include *Mbo* I, *Aci* I, *Bfa* I, *Dpn* II, *Hha* I, *Hin* P1I, *Hpa* II, *Mse* I, *Msp* I, *Nla* III, *Sau* 3AI, *Taq* I, *Csp* 6I, and *Tai* I.

Linker-ligated DNA fragments produced as described above are subsequently amplified, using an amplification reaction, such as a polymerase chain reaction, (U.S.

Pat. No. 4,683,202 by Mullis, K.B), nucleic acid sequence based amplification (NASBA) Kievits et al. (1991) J Virol Methods 35(3):273-286, ligation-mediated amplification (Volloch et al. (1994) Nucleic Acids Res 22(13):2507-2511, strand displacement amplification (SDA) (Walker et al. (1992) PNAC 89(1):392-396, sequence-independent single primer amplification (SISPA) (Reyes (1991) Mol Cell Probes 5(6):473-481, or ligase chain reaction (U.S. Pat. No. 5,686,272 issued to Marshall et al.

The support means provided in step (b) of the present method comprises a stationary support with at least one target oligonucleotide associated therewith. The stationary support preferably comprises a material capable of coupling with the oligonucleotide directly or indirectly. Suitable material capable of coupling directly with the oligonucleotide includes nitrocellulose, nylon, glass, silica, and latex. Examples of suitable stationary supports for use in this preferred embodiment of the present method include a nylon membrane, a filter embedded with silica particles, glass beads, silica magnetic particles, or a resin containing silica. Suitable material capable of coupling indirectly to the oligonucleotide through a first coupling agent bound to the oligonucleotide and a second coupling agent bound to the surface of the stationary support include avidin and streptavidin, or an antigen and antibody thereto.

The at least one target oligonucleotide associated with the stationary support includes a sequence of nucleotides which is complementary to a portion of the intermediate tandem repeat sequence of the DNA fragment. The term "portion" as used herein refers to a sequence of nucleotides within the ITR region of the DNA fragment of sufficient length that an oligonucleotide having a sequence complementary to the sequence would hybridize thereto when it comes into contact therewith. The "portion" is preferably a sequence of at least 20 bases in length, and more preferably a sequence of at least 40 bases. The target oligonucleotide more preferably has a sequence characterized by the formula $(A_wG_xT_yC_z)_n$, wherein A, G, T, and C represent the nucleotides which can be in any order; w, x, y and z represent the number of each nucleotide in the sequence and range from 0 to 7 with the sum of w+x+y+z ranging between 5 and 7; and n represents the number of times the sequence is tandemly repeated and is at least about 4 times, more preferably at least about 8 times, and most preferably at least about 15 times.

In step (c) of the method, the plurality of fragments of DNA is combined with

the support means under conditions wherein the DNA fragment containing the ITR hybridizes to the support means. When the plurality of fragments is a plurality of fragments of double-stranded DNA, the DNA is denatured prior to hybridization to the support means. Suitable means for denaturing double-stranded DNA fragments prior to hybridization to the support means include exposing the DNA to a temperature which is sufficiently high to denature double-stranded DNA, or suspension of the DNA in a denaturing solution. The DNA is preferably denatured using a denaturing solution containing a denaturing agent, such as a base (e.g. sodium hydroxide or potassium hydroxide). When a base is used to denature the DNA fragments, the pH of the resulting mixture is preferably adjusted to about a neutral pH, preferably by adding a buffer at a pH of about 4.8 to the mixture.

Once fragments of DNA have hybridized to the support means, the support means is preferably washed to remove DNA fragments and any other material present in any solution in which the support means is contained or on the surface of the support means which are not hybridized thereto. Any wash solution used is preferably configured to remove such materials without releasing the DNA fragments hybridized to the support means.

The DNA fragments hybridized to the support means can be released, from the support means using heat or an appropriate release solution, depending upon the nature of the association between the support means and the DNA fragments. For example, water or an aqueous low salt solution such as a TE buffer (e.g. 10 mM Tris-HCl, pH 7.5, 1 mM EDTA) can be used to release DNA fragments hybridized to a support means comprised of a silica material. Once released from the support means, the DNA fragments can be processed to further isolate DNA containing the ITR sequence from other fragments of DNA present in the resulting mixture of released DNA fragments. Additional processing steps could include rehybridization and screening according to the method described above, or cloning into a DNA vector and screening the transformants of the clones.

Figure 1 illustrates a preferred embodiment of the method of isolating a fragment of DNA containing an ITR, wherein a population of DNA fragments is prepared, hybridized to a support means, amplified, cloned, and screened for transformants containing the ITR. Each of the steps illustrated in Figure 1 is labeled with a roman numeral. Step I shows a molecule of double-stranded DNA (1) being

digested with a restriction enzyme (2), producing a population of DNA fragments (not shown) varying in size, at least one of which includes the target ITR. The arrow between Steps I and II illustrate a linker (3) being added to the population of DNA fragments to produce a population of linker-ligated fragments (8) with a linker (3) at the end of each of two different classes of DNA fragments, fragments with the target ITR sequence (6) and fragments without the target sequence (4). An oligonucleotide primer (7) having a sequence complementary to a priming sequence of each linker (3) is added to the population of DNA fragments (8) in Step III, and the population is amplified through a PCR reaction, thereby producing a population of amplified DNA fragments (9). In Step IV the population of amplified DNA fragments (9) is placed in a container (15) with a hybridization solution (12) and a filter (10) with at least one oligonucleotide having a sequence complementary to a portion of the target ITR sequence associated therewith. The hybridization solution promotes the hybridization of the DNA fragments containing the ITR sequence to the filter. In Step V, the filter (10) is removed from the container (15), and DNA fragments hybridized thereto are released therefrom. The resulting enriched population of released fragments are re-amplified in Step VI, using the same oligonucleotide primer (7) used in the amplification reaction in Step III. Finally, each fragment of the enriched amplified population of DNA fragments is cloned into a plasmid vector (18) in Step VII. The vectors are shown in Step VII cloned with fragments with the target ITR sequence (6) and cloned with fragments without the ITR sequence (4).

C. Method for Detecting a Polymorphic ITR Having Low Stutter:

Minimal stutter artifact is observed when a target ITR sequence of a DNA sample having such a sequence is detected according to this particular embodiment of the method of the present invention. The average stutter artifact observed is preferably no more than 1.1%, more preferably no more than 0.9%. The target ITR sequence can be either a perfect ITR or an imperfect ITR sequence. The DNA sample detected is preferably genomic DNA.

The average stutter artifact is preferably observed after amplification of the ITR sequence in the DNA sample.

D. Primers, Probes, and Markers

The present invention also comprises DNA markers identified in the Sequence Listing below as SEQ ID NO:'s 1-43, primers wherein each primer has a sequence which is complementary to a sequence flanking an ITR region of one of the DNA markers identified by one of those 43 sequences, and probes which have a sequence which is complementary to a sequence contained within the ITR region of one of the 43 markers. Specific preferred primers identified in experiments illustrated in the Examples, below are listed in Table 1.

TABLE 1

Marker SEQ ID NO	Clone Number	Primers SEQ ID NO	Upper Primer & Lower Primer
1	C074	44	TGGCTCAGACACCTCATTG
		45	CACCACTGTATTCCCAGTTTG
2	C221	46	CACTTGCCATCCCTGCCACACA
		47	AGCGCACCCCCAATTTCCGGTAT
	C221	48	TGGGGACATGAACACACTTTGC
		49	GAGGCCCAGGACCAGATGAAAT
	C221	50	CACCTGTCAGGCAAGGCTTAAAC
		51	CAACACTGAGCGCTTTTAGGGACT
	C221	52	TCAGGCAAGGCTTAAACAGGGATA
		53	ACACTGAGCGCTTCTAGGGACTTC
	C221	52	TCAGGCAAGGCTTAAACAGGGATA
		54	TGAGCGCTTCTAGGGACTTCTTCA
	C221	55	CCCTGCCCTACCCACTTG
		56	AGGCCCAGGACCAGATGA
	C221	57	GCACCTGTCAGGCAAGGCTTAAAC
		58	CCAGCCATGAAGTGGCTGTGAG
3	C240	59	CCCGCTTCAAAGTTCCCAGTTC
		60	CCTCCCATTTCAGCCTCCTGA
4	C331	61	GTCTGCCACAGTGCTGGAAACTAA
		62	GCACCCCAGCCTAAGGCAATA
5	C362	63	GCATGGCGGAAGAAACAA
		64	TGGCAACAGAGCGAGACTC
6	C390	65	CCTGGGTGACAGCGAGAATCT
		66	TGTCCCTTGCCTTGTCTCACTAAA
7	G022	67	CAGCCTTGGTGACAGAGCAAA
		68	TGTGTTGAGGGTGGGGTACAT
8	G023	69	CCTGGGCAAGAGAGCAAG

Marker SEQ ID NO	Clone Number	Primers SEQ ID NO	Upper Primer & Lower Primer
		70	CACATCCCAAAACCACCCTAC
9	G025	71	GCATTTCCCCTGCTTGTACT
		72	GATCACATTTGCTAACCACTTCTC
10	G047	73	GGCAACATATCAAGACCCCCATCTCT
		74	GAAGCTGCCCCTCACCACTACATTTT
11	G065	75	GATCACATTTGCTAACCACTTCTC
		76	TATAAATTACCCAGTCTCAGGAAG
12	G085	77	GTGATACAGCAAGCCTCATC
		78	AGAGACTCCTGGAAAGATAAAAGT
13	G132	79	GTCTGGAGAACAGTGGCCCTTGT
		80	CAGGAAGCTGAGGCAGGAGAATCT
14	G145	81	AAGGCTCCAGTGGGGTAT
		82	AAAACAAGGCAGTAGTCAATAAAG
15	G152	83	GGCATGAGAATCGCTTGAACCTG
		84	GGCCTCCATGATGTTTCCAATGAT
16	G153	85	TCAGGAGGCATGAGAATCGCTTGA
		86	GGCCTCCATGATGTTTCCCAATGA
17	G158	87	CTCGCCCTCTCCTATAAGCAGTTT
		88	GCAGAGATAATTTGGAGTGGGATG
18	G181	89	CTTGGGTGCCTGTAATCC
		90	GGTAGAGCTCCCCCATCT
19	G210	91	GCAGAATATTGGGGCTCATCAC
		92	AAACAAGGAAAGGAGAGGAGAGGA
	G210	93	AAGGTTGTGGGATGACTACTACA
		94	TGGTCAACACAGCAAGACATT
20	G212	95	TCCTGCCACCTGCTTGCTTTCT
		96	ATTGCACTCCAGCCTGGGTGATAC
21	G233	97	CGCTTGAGCCTTGGAGATTG
		98	GAGCAGTCAGAATTCAGGAGTTGT
22	G234	99	TGGGCAACAAGAGCAAAACTCCAT
		100	GGGACTTGGGCTGAGGGCTTTAC
23	G235	101	ATATCAATATCAGGCAGCCACAGG
		102	CCGTTTCAGAGCAGAGGTTTAGC
24	G331	103	TCTCATTGGTTTCAAAGAACTTA
		104	AGACTCCATCTCAAACAAAAGA
25	G405	105	TCATGTGCATGGAGCCTGGTTCAT
		106	CCCAGCCTTGGCAAGAGTGAGGT
26	G475	107	GGCGACTGAGCAAGACTC
		108	TTAAGCAAAGTAGCCTCAAACA
	G475	109	GGGCGACTGAGCAAGACTC
		110	ACTCATTACCTTGCATGCATGATA
	G475	107	GGCGACTGAGCAAGACTC

Marker SEQ ID NO	Clone Number	Primers SEQ ID NO	Upper Primer & Lower Primer
		111	CATTACCTTGCATGCATGATA
27	G539	112	TGGGCAACAGAGTAAGACTCA
		113	GTTCAGTACCGTTCACCTCTTTA
	G539	114	GTAAGACTCAGTCTCCAAAAAAAAAAAAAG
		115	AGGAATGGTTTCTCTGTTAGTAAATGGT
28	S023	116	CAGCCTGGGCAACAAGAATGAAAC
		117	TGGCCCCTGCAGCGGAGTC
29	S071	118	GAATTCATTTGCGGAAAGATT
		119	CTAGGGAGGCTGGAGTATTCA
30	S085	120	AGAGCAAGACCCCGTCTCAT
		121	AGTCCATGGGCCTTTTAACA
31	S125	122	GAGAATCACTTGAACCCAGGAAG
		123	AGAACCAGCTGTTAGTTTCGTTGA
32	S132	124	GGTTGCAGTGAGCCGAGATAAGAGT
		125	TGTGCCAGGAACCAGAAATTTACAG
33	S136	126	GGCCCAAGGTTACTTTTCAC
		127	GGGCCACTGCACTCCT
34	S159	128	CATGGTGAGGCTGAAGTAGGAT
		129	GTGGCGTGTCTTTTTACTTTCTTTA
35	S176	130	AGGCAGCCCAGGAACAAT
		131	CCAAGATAGCGGCCAAGATAGT
36	S189	132	GAGGGCAGCTGGGATGTTACTCTT
		133	TGCCCTGTTTGGAGAACTGTAGGT
37	S199	134	CTCCCCAGAAACAGATGTA
		135	GTGAGCCGAGATTGTATCAT
38	S040	136	TCGGGGACAGGGCTTACTC
		137	ATCATTGTCGCTGCTACTTTATCG_
39	S066	138	CTACTCTACCCCATTTCATTC
		139	GTAGAGTGGAGTGGATGAGA
40	S077	140	ATCAGGCAAAAACGAACAAAC
		141	CGGCATCCCAAAGTGAC
41	S097	142	CAGAGAGGGCAGCACCTTGGACAG
		143	GGCTTCACCTGCTCCCGTTTCAG
42	S103	144	TCTGCCCATTCCCCAGCCTCTC
		145	TACCGCGTGGCATTCAAGCATAGC
43	S110	146	TCCAGTCTGGGTGACAAA
		147	CAATCCACTCCACTCCTCTA

The following examples are offered by way of illustration, and are not intended to limit the invention in any manner. In the examples, all percentages are by weight

if for solids and by volume if for liquids, and all temperatures are in degrees Celsius unless otherwise noted.

Example 1 Construction of whole genome PCR library.

The particular amplification and hybridization selection techniques used in this Example, and in Example 2, below, are modified forms of a selection method described in Armor, J. et al. (1994) *Hum Mol Genet* 3(4):599-605.

Human genomic DNA was purified from whole blood pooled from 15 individuals using standard phenol:chloroform extraction procedures (Current Protocols in Human Genetics (1994), Gilber, J. ed., Appendix).

Approximately 100 µg genomic DNA was cut with 5 units of *Mbo* I restriction enzyme per µg of DNA for 16 hrs at 37°C, followed by purification with by phenol:chloroform extraction, ethanol precipitation and resuspended in 100 µl of TE Buffer (10mM Tris-HCl, 1mM EDTA, pH 8.0) for a final concentration of about 1 µg/µl of DNA.

DNA fragments ranging in size from 250-600 bp were isolated by gel electrophoresis on a 1% SeaKem GTG (FMC Bio Products, Rockland, Maine) preparative agarose gel (15x20 cm) for 1.25 hours at 100 volts and recovered by electroelution (reference). The DNA was quantified by measuring absorbance at A_{260} and diluted to 500 ng/µl in sterile nanopure water and stored at -20°C .

Linkers were prepared by annealing equimolar amounts of oligo A (5'-GCG GTA CCC GGG AAG CTT GG-3') and 5' phosphorylated oligo B (5'-GAT CCC AAG CTT CCC GGG TAC CGC-3') for a final concentration of 1,000 pmol/µl. One µg of size selected insert DNA (3.5 pmols with an average size of 425bp) was ligated to 13 µg (875 pmols) of linkers (250:1 linker:insert molar ratio), using 1-3 units of T4 DNA ligase for 16 hr at 15°C. Excess linkers and linker dimers were separated from the primary fragments by gel electrophoresis (1% SeaKem GTG agarose, 1.5 hrs at 100 volts). The linker-ligated DNA fragments were recovered from the gel by electroelution, and resuspend in 50 µl sterile water.

DNA (50ng) with ligated linkers were amplified using a PCR in 100 µl reaction volume containing 10µl of a 10X STR buffer (500 mM KCl, 100 mM Tris-HCl, pH 9.0, 15 mM $MgCl_2$, 1% Trition X-100, and 2 mM of each dNTP), 1 µl *Taq* polymerase (5U/µl), and 1 µM oligo A primer (10 µl of a 10 pmol/µl stock). The "oligo A" used as

a primer in this reaction is the same "oligo A" used to assemble the *Mbo* I linker, as described above. Cycling conditions were 95°C 1 min, 67°C 1 min, 70°C 2 min; for 30 cycles. The dNTPs, primers and primer dimers were removed by microfiltration with Centricon-100s (add 2 ml sterile water to sample and load Centricon-100, spin 20 min at 2,000 RPM, invert Centricon filter and spin for 2 min at 2,000 RPM to recover DNA, resuspend in 100 µl sterile dH_2O). A 5 µl aliquot of the resulting PCR library was checked on 1% agarose gel (1hr at 100 volts) to confirm that the size range was between 250 and 600 bp.

Example 2 Enrichment for pentanucleotide repeats by hybridization selection.

DNA fragments from the whole genome PCR library produced according to Example 1 containing various different repeats were enriched by hybridization using different oligonucleotide mixtures associated with a solid support. Fragments containing $(AAAAX)_n$ pentanucleotide repeats were enriched by hybridization selection. This process was accomplished by first constructing oligonucleotides for use in hybridization selection that consisted of tandem arrays of $(AAAAC)_n$, $(AAAAG)_n$ and $(AAAAT)_n$ around 1000 bp in length. These oligonucleotides were fixed to membranes and hybridized to the whole genome PCR library to select those fragments containing $(AAAAX)_n$ repeats.

The array of oligonucleotides was constructed as follows: (a) 5'-phosphorylated 30 mer oligonucleotides of $[AAAAC]_6$, $[AAAAG]_6$ and $[AAAAT]_6$ and their complements $[GTTTT]_6$, $[CTTTT]_6$ and $[ATTTT]_6$ were synthesized and suspended in nanopure water at a concentration of 1,000 pmol/µl, (b) equal molar concentration (used 10 µl or 10 nmol or 198 µg each) of oligonucleotides having complementary sequences were combined, heated to 65°C for 15 minutes and left at 4°C for a few hours to anneal to one another, (c) the annealed oligonucleotides were then ligated to one another using 1 Weiss Unit of T4 DNA ligase per µg DNA at 15°C overnight, (d) concantomers ≥200 bp were size-selected on 1% SeaKem GTG agarose, (e) the ligated DNA was subjected to primer-free PCR to lengthen the tandem arrays, (f) fragments of apparent size over 1000bp were recovered from 1% agarose gels and purified by microfiltration. The absorbance at A_{260} was determined and a one µg/µl stock was made in sterile nanopure water.

A total of one µg of $(AAAAC)_{200}$, $(AAAAG)_{200}$, or $(AAAAT)_{200}$ oligonucleotide was then spotted onto 4mm x 4mm pieces of nylon Hybond-Nfp membrane (Amersham Life Sciences, Inc.) filter, washed twice in pre-hybridization buffer for 30 minutes with agitation to remove weakly bounded oligos, allowed to air dry, UV cross-linked at 1200 µJoules to bind DNA, then stored at -20°C.

Hybridization selection of the whole genome PCR library to the resulting support medium of oligonucleotides associated with the nylon filter described above was accomplished as follows: (a) the filters were prehybridized in 1 ml Prehybridization Buffer [1% BSA (Sigma B-4287), 1mM EDTA, pH 8.0, 7% (w/v) SDS, 0.5M Na_2HPO_4] at 40°C for filters containing oligonucleotides having sequences of $(AAAAC)_n$ and $(AAAAG)_n$ and at 37°C for those containing $(AAAAT)_n$ sequences. After 20 minutes the buffer is removed and 100 µl of fresh Prehybridization Buffer is added, (b) whole Genome PCR Library DNA (20 µg) was denatured with alkali (KOH, final concentration 150mM) and neutralized by adding 0.25 volumes of 1M Tris-HCl pH 4.8 and added to the buffer containing the filters. The resulting reaction mixture was incubated overnight at prehybridization temperatures of 37°C or 40°C, (c) the $(AAAAC)_{200}$ and $(AAAAG)_{200}$ filters are washed 2X with 1 ml Wash Buffer #1 (40mM Na2HPO4, pH 7.2, 0.1% SDS) at 40°C and 1X at room temperature for 15 minutes with agitation. The $(AAAAT)_{200}$ filters are washed 1X with 1 ml Wash Buffer #1 at 37°C and 1X at room temperature, (d) DNA bound to each filter was released by heating to 95°C for 5 minutes in 100 µl sterile nanopure water. The sample was removed while at 95°C to prevent re-annealing. Filters were stripped and reused by incubating in 0.4M NaOH for 30 minutes at 45°C, then transferring to 0.1X SSC, 0.1% SDS, 0.2M Tris pH 7.5 and incubating another 15 minutes. The membranes were blotted dry and stored in sealed tubes at -20°C.

Example 3 Cloning pentanucleotide repeat enriched library of DNA fragments.

The population of DNA fragments enriched for pentanucleotide repeats according to Example 2 was re-amplified by PCR. The reamplified fragments were then cloned into plasmid vector pGEM-3Zf(+), as described below. This process was accomplished by ligating selected inserts to the pGEM vector then transforming circularized plasmid into a JM109 *E. coli* host.

The insert-vector ligations were accomplished as follows: (a) 5 µl of the hybridization selected DNA was reamplified in a 100 µl reaction volume, using a 1XSTR buffer (50 mM KCl, 10mM Tris-HCl, pH 9.0, 1.5 mM $MgCl_2$, 0.1% Triton X-100, and 0.2mM each dNTP), 1 µl *Taq* polymerase (5U/µl), and 1 µM oligo A primer (1 µl of 100 pmol/µl stock). Cycling conditions were 95°C 1 min, 67°C 1 min, 70°C 2 min; for 30 cycles. (b) The reamplified DNA was digested with *Mbo* I by adding 11µl Promega restriction enzyme 10X Buffer C and 2µl (8U/µl) *Mbo* I to the 100 µl PCR reaction, by incubating the resulting reaction mixture overnight at 37°C, and by heat inactivating the restriction enzyme by incubating the mixture at 65°C for 20 minutes. (c) The pGEM-3Zf(+) vector (~20 µg or 10.6 pmol) was prepared for fragment insertion by digesting with *BamH* I (5U/µg) for 16 hours at 37°C, followed by the addition of appropriate amounts of Calf Intestinal Alkaline Phosphate 10X buffer (Promega) and 1 µl CIAP (Units/µl) and incubation for 1 hour at 37°C. This reaction was stopped by adding 0.5M EDTA to 0.02M final concentration then phenol extracted, ethanol precipitated and resuspend in TE buffer at 1µg/µl. (d) Finally, 20 µl insert-vector ligations were performed by incubating 1 µl of DNA cut with *Mbo*I (see step b) along with 1 µl or 200ng of dephosphorylated pGEM 3Zf(+) (see step c) and 1 µl T4 DNA ligase (1 to 3 U/µl) for 2 hours at room temperature.

Finally, 10 µl of the insert-vector ligation reaction were transformed into 100 µl of JM109 competent cells using the Promega transformation protocol described in Technical Bulletin #095.

Example 4 Selection of small insert genomic library clones containing $(AAAAX)_n$ pentanucleotide repeats by colony hybridization.

Clones containing $(AAAAX)_n$ pentanucleotide repeats were selected by colony hybridization screening using Lightsmith II reagents and protocols (see Promega Technical Bulletin #TM227), and visualized by hybridization to alkaline phosphatase conjugated probes.

Colony DNA was transferred to membranes by placing MagnaGraph nylon membranes (Micron Separations, Inc. Westboro, MA) on plates containing bacterial colonies, allowed to sit for 3 minutes, then blotting on dry filter paper. Next, the membranes were transferred to a series of trays containing 10% SDS for 3 minutes, then denaturing solution consisting of 5ml NaOH + 30ml 5M NaCl + 65ml dH_2O for

5 minutes, then Neutralizing solution consisting of 30ml 5M NaCl + 25ml M Tris-HCl, pH 7.4 + 45ml dH_2O for 5 minutes, and finally 2X SSC for 5 minutes. The membranes were then dried at room temperature for 30 minutes followed by UV crosslinking with 1200 µjoules, using a Statalinker® (Stratagene, La Jolla, CA).

Detection of colonies containing clones with $(AAAAX)_n$ repeats was accomplished with the aid of AP conjugated probes and chemiluminescence. Exposure of filters hybridized to AP conjugated probes to X-ray film indicated colonies contain desired clones. A second hybridization was performed to confirm initial results.

The detection procedure utilized Lightsmith II kit from Promega (see Promega Bulletin #TM227 for detailed description of the procedure). Briefly, the detection procedure used consisted of the steps of: (a) Incubating of the filters in a Quantum Yield® Blocking Solution (Promega Cat NO F1021) for 45 minutes at 56°C with vigorous shaking, (b) pouring off the Blocking Solution and adding 0.05 ml of Quantum Yield® High Stringency Hybridization Solution (Promega Cat No. F1231) per cm^2 of membrane containing the AP probe and incubating 45 minutes at 56°C with vigorous shaking, (c) pouring off the hybridization/probe solution from the filters and wash filters twice with 150-200 ml of preheated Wash Solution #1 (2X SSC, 0.1% SDS) for 10 minutes at 56°C, (e) combining all filters and wash once with Wash Solution #2 (1X SSC) for 10 minutes at room temperature, (f) equilibrating the blots for 5 minutes in 200 ml of 100mM diethanolamine, 1mM $MgCl_2$, (f) adding sufficient 0.25mM CDP-Star substrate (Tropix, Bedford, MA) to saturate filters then incubate for at least 5 minutes at room temperature, (g) placing the substrate-saturated filters on a polystyrene plastic sheet protector in a hybridization folder and closing the folder, (h) placing the hybridization folder containing the filters in a film cassette and exposing the filters contained therein to X-ray film, and (I) developing the film after at least a 1 hour period of exposure to the film.

Example 5 DNA sequencing and analysis.

A simplified method of preparing sequencing templates utilizing cell lysates was developed to sequence the large number of clones identified in Example 4 as possibly containing inserts with at least one $(AAAAX)_n$ sequence. This procedure consisted of transferring positive clones from colony hybridization assays to sterile

96 well microtiter plates (Falcon cat. # 3072) containing 200 µl of LB/Amp (100µg/ml) and incubating overnight at 37°C at 250 rpm. Next, the overnight culture was divided and used in three different procedures involving either setting up of the cell lysates, making replica filters for second hybridizations to confirm initial findings or making glycerol stocks for long term storage of clones.

Cell lysates were made by taking 2µl of overnight culture and adding this to 100µl sterile nanopure water in 96 well reaction plates (Perkin Elmer cat. # N801-0560) and heating to 100°C for 4 minutes in 9600 thermocycler. This was allow to cool, iced, and stored at -20°C until ready to use.

Replicate filters were made for second hybridization assays by flame sterilizing the 96-pin replicator, dipping the replicator into a 96 well plate containing overnight culture and stamping a 137 mm circular nylon membrane (MagnaGraph, MSI) on a LB / Amp (100 µg/ml) plate and incubating the membrane overnight at 37°C.

The remaining overnight culture was converted to glycerol stocks by the addition of 46µl 80% glycerol to each well and placing plates on in shaker -incubator set on 250 rpm for a few minutes to mix, then stored at -70°C.

All clones that were positive in two colony hybridization assays were selected and corresponding clones from the cell lysate plates were used for PCR amplification. The PCR reaction products were purified with Qiagen QIAquick 96 PCR Purification plates (Cat. #28180) and used a templates for sequencing. Two microliters of the cell lysate were used in a 50 µl PCR reaction containing M13 -47 forward primer at 2 µM (Promega cat. #Q560A), M13 reverse primer (Promega cat. #Q542A) at 2 µM, 1X STR buffer and 2.5 units of AmpliTaq (Perkin Elmer). The following cycle profile was used on a PE 480 thermocycler: 1 cycle at 96°C / 2 min, 10 cycles at 94°C / 1 min, 56°C / 1 min, 70°C / 1.5 min; 20 cycles at 90°C / 1 min, 56°C / 1 min, 70°C / 1.5 min; 4°C hold. PCR reaction products were clean-up with Qiagen QIAquick 96 PCR Purification plates (Cat. #28180) following manufacturers protocol and recovered in 70µl Tris-HCl 10mM pH 8.5 at a final concentration of about 35 ng/µl and stored at -20°C.

DNA sequencing was performed using ABI Dye Terminator Sequencing Chemistry and ABI 377 sequencer. The sequencing templates were prepared using ABI Dye Terminator Kit and manufactures protocol (Protocol P/N 402078). Two µl or approximately 30 to 90 ng of purified PCR product (described above) was used

as a template DNA for sequencing reaction. The sequencing reaction consisted of 8 µl Dye terminator mix, 2 µl template DNA (35ng/µl), 4 µl of M13 -21 Forward primer at 0.8 µM, and 6µl of sterile nanopure water. Cycle sequencing on the GeneAmp PCR System 9600 cycling profile was: 25 cycles at 96°C / 10 sec, 50°C / 5 sec, 60°C / 4 minutes; hold 4°C. The extension products were purified by adding 50µl 95% ethanol and 2µl 3M Sodium acetate, pH 4.6 to each tube, mixed using a vortexer, placed on ice for 10 minutes, then centrifuged for 30 minutes at maximum speed. The pellet was rinsed with 250µl 70% ethanol, dried in vacuum centrifuge for about 3 minutes and stored dry at -20°C until ready for use. The dried pellet was resuspended in 6-9µl loading buffer then denatured for 2 minutes at 95°C and stored on ice until loaded on gel.

Five percent Long Ranger gels (FMC BioProducts, Rockland, ME) were prepared according to manufacturer protocol and polymerized for 2 hours. The gel was pre-run for 45 minutes at 1000 volts. 1.5 µl template in loading buffer was loaded on gel and run under 2X or 4X conditions for 3.5 hrs or 7 hrs, respectively.

DNA sequence data generated from the ABI 377 sequencer was edited to remove any pGEM vector sequences then placed in local database created using Genetics Computer Group Wisconsin Package Software version 9.0 (Madison, WI) containing sequence information for all clones being evaluated. Next, clones were examined for the presence, length and sequence patterns of pentamer repeats. Those containing 5 or more repeats were then compared with the BLAST sequence comparison program (Altschul et. al., 1990) to identify duplicated clones and those that already existed in GenBank database at the National Center for Biotechnology Information in Besthesda, Maryland, USA. Once unique clones were identified, primers were designed for PCR with the aid of OLIGO Primer Analysis Software version 5.0 (National Biosciences, Inc., Plymouth, MN).

Example 6 Screening clones for polymorphism levels and determining chromosomal location.

The Initial screen for polymorphisms was performed on two pooled DNA samples, one containing human genomic DNA 15 random individuals and the other containing 54 CEPH individuals from the NIGMS Human Genetic Mutant Cell Repository (CEPH Collection DNA Pool, cat. #NA13421, Coriell Cell Repositories,

Camden, NJ). Fluorescently labeled PCR primers were used for PCR amplification of target locus from genomic DNA and the PCR products were separated on polyacrylamide gels and visualized on a fluorescent scanner. Those loci with 4 alleles and 50% heterozygosity were subsequently tested with 16 individual CEPH DNAs (102-1, 102-2, 884-1, 884-2, 1331-1, 1331-2, 1332-1, 1332-2, 1347-1, 1347-2, 1362-1, 1362-2, 1413-1, 1413-2, 1416-1, 1416-2) to determine preliminary heterozygosity values. The data for the same loci was then further analyzed to determine number of alleles, allele frequencies and heterozygosity values (see TABLE 2).

Clones found to contain pentamer repeat sequences that met the selection criteria of ≥4 alleles and ≥50% heterozygosity were mapped to determine precise chromosomal location (see TABLE 2). Three different methods were used for mapping: (1) Somatic cell hybrid mapping using the NIGMS panel of 26 somatic cell hybrids (Coriell Cell Repositories, Camden, NJ) representing single human chromosomes to identify chromosomal origin, (2) radiation hybrid mapping techniques utilizing the GeneBridge 4 RH Panel of 93 RH clones (Schuler et. al., 1996), and (3) standard meiotic linkage mapping techniques and eight families (K102, K884, K1347, 1362, 1331, 1332, 1413, 1416) from the CEPH kindred reference panel and mapped with CRI-MAP multipoint linkage program (Lander & Green, 1987).

Clones with heterozygosity values exceeding 70% in the 16 CEPH individuals were evaluated for genotype and allele frequencies in larger population studies containing over 100 individuals from four major races, including, African Americans, Caucasians, Asians, and Hispanics. Figures 10 and 11 illustrate the wide variation in the migration of alleles amplified from two different polymorphic ITR loci in genomic DNA samples from 24 different individuals in a population (DNA samples S02 to S25). See Table 1, above, for the sequence of the primer pairs used in this analysis. The gel images were generated by amplifying each pentanucleotide repeat locus using fluorescein labeled primers, followed by separation on polyacrylamide gels and visualized by scanning of the FMBIO II Fluorescent Scanner (Hitachi Software Engineering America, Ltd., San Francisco, CA). An alleleic ladder containing most known alleles for each locus assayed was included in a lane at each end of the electrophoresis gel, in lanes S01 and S26. The primer pairs used to amplify each

locus had sequences complementary to at least a portion of the sequence of a DNA marker isolated from clone S159 or from clone G210, as illustrated in the Examples above. The primer pair sequences were selected from the primer pairs listed for Clones S159 and G210 Table 1, above.

PCR conditions for polymorphism screens were as follows: 25µl reactions containing approximately 200ng for pooled DNA template or 25ng for individual CEPH DNAs, 1X STR Buffer, 1 unit *Taq* DNA Polymerase, and 1µM corresponding primer pair. The sequence of each primer pair used to amplify each of the clones listed in Table 2 is provided in Table 1. Note that each primer has been assigned the SEQ ID NO listed in Table 1. Cycling conditions for the Perkin-Elmer GeneAmp PCR System 9600 Thermal Cycler (Perkin-Elmer, Foster City, CA) were: 96°C for 1 minute, then 10 cycles at 94°C for 30 seconds, ramp 68 seconds to 60°C, hold 30 seconds, ramp 50 seconds to 70°C, hold for 45 seconds; followed by 20 cycles of 90°C for 30 seconds, ramp 60 seconds to 60°C, hold for 30 seconds, ramp 50 seconds to 70°C, hold 45 seconds, 60°C for 30 minutes. PCR Samples were prepared by mixing 2.5µl of each sample with 2.5µl 2X Bromophenol Blue Loading Solution, denatured by heating at 95°C for 2 minutes, iced, then 3µl of each sample was run on a 4% polyacrylamide gel for 50 minutes at 40 watts. The PCR products were visualized by scanning of a Hitachi FMBIO fluorescent scanner and analyzed with accompanying software (FMBIO Analysis Version 6.0, Hitachi Software Engineering, San Francisco, CA).

TABLE 2

SEQ ID NO.	Clone Number	GenBank Accession Number	Longest ITR Sequence Observed	Observed No. of Alleles	% Hetero-zygosity (Caucasians)	Chromosomal Location
1	C074	none	$[TTTTG]_9$	6	75	1
2	C221	none	$[GTTTT]_{13}$	7	78	9p
3	C240	none	$[CAAAA]_7$	4	42	NA
4	C331	none	$[GTTTT]_{10}$	5	43	NA
5	C362	none	$[GTTTT]_5$	4	62	4
6	C390	none	$[CAAAA]_7$	5	56	NA
7	G022	none	$[AAAAG]_6$	4	63	2p
8	G023	none	$[AAAAG]_{10}$	12	71	16q
9	G025	none	$[AAAAG]_6$	12	86	1
10	G047	none	$[AAAAG]_9$	5	86	2p
11	G065	none	$[TTTTC]_6$	13	100	1q

SEQ ID NO.	Clone Number	GenBank Accession Number	Longest ITR Sequence Observed	Observed No. of Alleles	% Hetero-zygosity (Caucasians)	Chromosomal Location
12	G085	none	$[AAAAG]_{11}$	8	93	10q
13	G132	none	$[CTTTT]_{15}$	12	100	4 qter
14	G145	none	$[AAAAG]_{13}$	8	33	NA
15	G152	none	$[AAAAG]_{6}$	5	87	8 qter
16	G153	none	$[AAAAG]_{6}$	5	88	8 qter
17	G158	none	$[AAAAG]_{5}$	8	75	5q
18	G181	none	$[GAAAA]_{14}$	5	72	NA
19	G210	none	$[CTTTT]_{6}$	9	56	8p
20	G212	none	$[CTTTT]_{9}$	6	100	NA
21	G233	none	$[AAAAG]_{8}$	12	50	10q
22	G234	none	$[AAAAG]_{12}$	4	80	16 qter
23	G235	none	$[TTTTC]_{6}$	4	56	2p
24	G331	none	$[CTTTT]_{8}$	5	73	NA
25	G405	none	$[CTTTT]_{6}$	10	80	NA
26	G475	none	$[GAAAA]_{12}$	12	92	15q22.3
27	G539	none	$[GAAAA]_{12}$	13	100	15q26.2
28	S023	X05367	$[AAAAT]_{6}$	4	50	NA
29	S071	M90078	$[AAAAT]_{8}$	4	56	6q26-27
30	S085	U07000	$[AAAAT]_{5}$	7	44	22q11
31	S125	Z73416	$[AAAAT]_{13}$	5	64	22q11.2-qter
32	S132	Z83847	$[AAAAT]_{10}$	8	69	22
33	S136	Z82250	$[TTTTC]_{6}$	11	94	22q12-qter
34	S159	AC000014	$[GAAAA]_{9}$	12	72	21q22-qter
35	S176	AC000059	$[GTTTT]_{9}$	4	56	7q21-7q22
36	S189	Z54073	$[AAAAC]_{8}$	5	69	22q11.2-qter
37	S199	Z84475	$[GTTTT]_{7}$	4	75	6q21
38	S040	X06583	$[AGCCTGG]_{4}$	2	NA	NA
39	S066	M68516	$[ACTCC]_{5}$	3	NA	NA
40	S077	M25718	$[[AATAC]_{12}$	6	NA	NA
41	S097	Z21818	$[CAGGCT]_{3}$	3	NA	NA
42	S103	X15949	$[ATCCC]_{8}$	3	NA	NA
43	S110	X54108	$[GGA(A/G)T]_{32}$	6	NA	NA

Example 7 Identification of short tandem repeats through GenBank searches.

An alternate method of identifying tandemly repeated sequences was accomplished by searching GenBank at the National Center for Biotechnology Information (NCBI) for the presence of intermediate tandem repeats. Several methods were employed, including batch searching of GenBank entries on CD-ROM with the Lasergene software package from DNASTAR (Madison, WI), batch searching GenBank with the aid of Genetics Computer Group Wisconsin Package Software version 9.0 (Madison, WI).

There are 4^5=1024 distinct five letter words which can be assembled from the four letter (A, C, G, and T) alphabet to make all the possible pentamer repeats, and

4^6=4096 and 4^7= 16,384 distinct six and seven letter words for six and seven base repeats. However, the number of unique repeat motifs is considerable less due the equivalence of the two complementary strands (e.g., AAAAT is equivalent ATTTT), to and the equivalence of cyclic permutations (e.g., AATAA. . . is equivalent to ATAAA. . .). In the case of five base repeats, this means that there exists 102 unique classes of pentamer repeats if one leaves out mononucleotide repeats A_5/T_5 and C_5/G_5.

All unique combinations of 5, 6 and 7 base repeats with at least three consecutive copies were used to search the GenBank human genome database. All repeat regions containing three or more copies of a repeat, or copies with occasional base substitutions, were identified. Using existing sequence data, primers flanking the repeat region were designed and the target locus was PCR amplified and evaluated for polymorphic content as described in Example 6.

Each clone containing a sequence identified using primers assembled using information from the GenBank database was then screened for repeat sequence content as described in Example 7. The sequence of each clone found to contain an ITR sequence, i.e. an ITR marker, was assigned one of the SEQ ID NO's from 28 to 43. See Table 1 for the sequence of primers comprising sequences which flank the ITR region of each such marker. See Table 2 for a summary of results of analyzing the characteristics of the sequence of each such ITR marker.

Example 8 Evaluation of Intermediate Tandem Repeat loci for PCR artifacts (i.e., % stutter).

Many of the markers described in this work represent a new class of markers which produce less of a PCR artifacts known as "stutter" (see Definitions section of the Detailed Description of the Invention, above). The generation of these artifacts occurs during PCR amplification, presumably as a result of a DNA polymerase-related phenomenon called repeat slippage (Levinson & Gutman, 1987. Mol. Biol. Evol. 4(3):203-221; Schlotterer & Tautz, 1992. NAR 20:211-215). The end result of repeat slippage is the generation of PCR products that contain different numbers of repeat units than the authentic allele. If sufficient amount of slippage occurs during PCR, the amplified product will be visualized as a major and minor band, with the major band corresponding to the authentic allele and the minor band corresponding

to the altered product containing more or less of the repeat units.

To quantify the amount of the stutter band present at different loci, PCR amplification products of 6 ITR loci (C221, GO23, G025, G210, S159 and an additional ITR not described in this patent, S117) and 17 tetranucleotide tandem repeat loci (F13A01, THO1, TPOX, F13B, FESFPS, D7S820, CSF1PO, D13S317, D8S1179, D16S539, LPL, FGA, D5S818, D3S1358, D18S51, vWA, and D21S11) were run on an ABI 377 Sequencer and analyzed using GenScan software (PE Applied Biosystems, Foster City, CA). The peak heights measured in relative fluorescence units (RFU) were determined for all major and minor peaks observed in the 25 to 40 individual samples investigated at each loci. The percentage of RFU observed in the minor peak (generally either 5 bp smaller than the authentic allele in the pentanucleotides or 4 bp smaller in tetranucleotide repeats) to the major authentic allele peak was calculated (see Table 3).

Examples of ABI 377 electropherograms for ITR loci S159 (Fig. 2) and G210 (Fig. 3) and tetranucleotide repeat loci vWA (Fig. 4) and D5S818 (Fig. 5) show minimal or absent stutter at ITR loci and clearly observable stutter for tetranucleotide repeat loci. Specifically, see the stutter artifacts indicated by arrows 14 and 15 in the electropherogram of the vWA tetranucleotide repeat locus reproduced in Figure 3, and by arrows 16 and 17 in the electropherogram of the D5S818 tetranucleotide repeat locus reproduced in Figure 5. Compare those distinct artifact peaks to the vanishingly small artifacts in electropherograms of the pentanucleotide repeats of the marker DNA isolated from Clone S159 (i.e. marker having the sequence identified by SEQ ID NO:34) as shown in Figure 2, and of the marker DNA isolated from Clone G210 (i.e. marker having the sequence identified by SEQ ID NO:19) in Figure 4. The specific electropherograms reproduced in Figures 2 - 5 are the highest incidences of stutter observed for each of the loci.

Some variability in the amount of stutter was observed for all loci. In general the trend was for alleles containing the highest number of repeats (as indicated by their size in base pairs) to exhibit the highest amount of stutter. Percent stutter values for each of the 25 to 40 individuals tested are shown is scatter plots (figures 6, 7, 8 and 9).

In summary, the percentage of the "stutter" band to the authentic allele band was significantly lower in most of the ITR loci evaluated compared to the

tetranucleotide tandem repeat loci. This was true even though the tetranucleotide loci used represent the best of this type of marker currently known. For example, 13 such tetranucleotide markers, including several of the tetranucleotide markers assayed as described reported in Table 3 below as having a high % stutter, have been selected by the U.S. Federal Bureau of Investigation for use in analyzing all DNA samples for the national Combined DNA Index System (CODIS). (Macivee, I. (1998) *Profiles in DNA* 1(3):2).

TABLE 3

Locus Name or Clone Number	Tandem Repeat Unit Length	Average Percent Stutter	Highest Percent Stutter	Lowest Percent Stutter	Standard Deviation	Number of Alleles Analyzed
Clone S159	5 bp (ITR)	0.1	1.4	0.0	0.4	40.0
Clone G210	5 bp (ITR)	0.6	3.2	0.0	0.9	30.0
Clone C221	5 bp (ITR)	0.9	3.3	0.0	0.9	27.0
F13A01	4 bp	1.2	9.7	0.0	2.5	34.0
TH01	4 bp	1.7	5.2	0.0	1.7	34.0
Clone S117	5 bp (ITR)	2.0	6.9	0.0	1.7	37.0
Clone G023	5 bp (ITR)	2.3	6.6	0.0	1.7	39.0
TPOX	4 bp	2.4	5.6	0.0	1.8	34.0
F13B	4 bp	2.6	7.7	0.0	1.7	31.0
FESFPS	4 bp	3.6	10.0	0.0	2.3	34.0
D7S820	4 bp	3.8	8.2	1.6	1.6	28.0
CSF1PO	4 bp	4.1	9.5	0.0	2.5	31.0
Clone G025	5 bp (ITR)	4.5	9.3	0.0	2.1	36.0
D13S317	4 bp	4.7	7.5	1.7	1.5	26.0
D8S1179	4 bp	5.0	8.3	2.4	1.6	27.0
D16S539	4 bp	5.1	8.6	1.7	2.0	28.0
LPL	4 bp	5.4	15.0	1.7	3.1	29.0
FGA	4 bp	5.5	11.6	3.0	1.7	36.0
D5S818	4 bp	6.1	9.0	0.0	1.9	28.0
D3S1358	4 bp	6.1	12.5	0.9	2.1	25.0
D18S51	4 bp	6.5	11.6	2.5	2.4	28.0
vWA	4 bp	6.6	11.4	3.7	1.4	28.0
D21S11	4 bp	7.5	15.7	1.9	3.5	30.0